ifs

WORKING PAPER SERIES

NO. 88/15

DYNAMIC PANEL DATA ESTIMATION USING DPD - A GUIDE FOR USERS

MANUEL ARELLANO
STEPHEN BOND

THE INSTITUTE FOR FISCAL STUDIES
180/182 TOTTENHAM COURT ROAD
LONDON W1P 9LE
TELEPHONE: 01–636 3784

DYNAMIC PANEL DATA ESTIMATION USING DPD

– A GUIDE FOR USERS

by

Manuel Arellano* and Stephen Bond**

September 1988

*Institute of Economics and Statistics
Manor Road, Oxford, OX1 3UL

**Institute for Fiscal Studies
180/182 Tottenham Court Road, London, W1P 9LE.

DPD was originally developed to use with the IFS company database as part of the IFS project on corporate behaviour and the impact of taxation. We have benefited from the input of many colleagues, but would particularly like to thank Richard Blundell for his encouragement and many helpful suggestions. This research was supported by ESRC grant B00232207.

INTRODUCTION

DPD is a program written in the Gauss matrix programming language to compute estimates for dynamic models from panel data. A number of estimators are available, including the generalised method of moments (GMM) technique developed in Arellano and Bond (1988), as well as more familiar OLS, within-groups and instrumental variables procedures. Standard errors and test statistics that are robust to the presence of heteroskedasticity are provided. Tests for serial correlation and instrument validity are automatically computed. Further tests of linear restrictions and sub-sample stability are available as options. Lagged and differenced series are easily constructed, with many other data transformations available. A particularly attractive feature of DPD is that it allows estimates to be computed from panels that are unbalanced in the sense of having a variable number of time-series observations per individual unit. In many contexts this allows a much larger sample to be exploited than would be the case if a balanced panel were required.

DPD was developed to use with data on a panel of companies, but is applicable to many other situations in which the number of time-series observations is small and the number of cross-section observations is large. We concentrate on estimators that do not require regressors to be strictly exogenous, and which require only the cross-section dimension of the data set to become large for consistency.

Section 1 of this guide describes how DPD can be installed and how data should be organised for use with DPD. Section 2 contains an account of the econometric methods employed by DPD. Section 3 provides detailed instructions on how to use DPD, and Section 4 contains an example.

1. INSTALLATION

DPD is contained in 3 files: DPD.RUN, DPD.FNS and DPD.PRG. These were written using Version 1.49B of the Gauss language, but will certainly run with earlier versions. DPD will run on an XT with maths co-processor, but for speed reasons it is recommended that an AT (or faster) microcomputer is used. For the same reason the program should only be run from hard disk.

Where Gauss has been fully installed the DPD files should be copied either on to the \Gauss subdirectory or on to the subdirectory where you keep the rest of your programs – either is acceptable provided the three DPD files are all in the same location. Note that since DPD does not require any external functions, it is not even necessary for Gauss to have been properly installed. Where this applies simply copy the DPD files to the drive and subdirectory where GAUS*.* are located.

1.1 DATA

The main requirement for running DPD is a suitably ordered Gauss data set. This will normally be created from a sorted ASCII data file, using the dos atog command (see the Gauss manual for more details). Each row should contain data on a set of variables for some cross-sectional unit and some time period. Each column should refer to the same variable in every row, and one column must contain the year to which the observation refers (in the form 19xx). All observations for each individual unit must be consecutive and together sequentially in the data set. Individual units on which there is a common number of time-series observations should be grouped together. It is sometimes useful, although not essential, for these groups of cross-sectional units to appear in ascending (or descending) order of the number of observations per unit.

In addition to this main Gauss data set, DPD requires a secondary or auxiliary Gauss data set which describes the structure of the main data. This auxiliary data set must contain two columns: elements of the first contain the number of observations per individual unit in the appropriate section of the main data file; and elements of the second contain the number of individual units which have this number of observations. This structured description begins from the top of the main data set and moves down it. For example, if the main data set contains information on 40 companies, with 8 obervations on each of the first 20 firms and 10 observations on each of the second 20, then the auxiliary data set will take the form:

8	20
10	20

Again this will normally be created from an ASCII file using the dos atog command. It is used to facilitate the reading of unbalanced data sets.

2. ECONOMETRIC METHODS

The general model that can be estimated with DPD is a single equation with individual effects of the form:

$$y_{it} = \sum_{k=1}^{p} \alpha_k y_{i(t-k)} + \beta'(L) x_{it} + \lambda_t + \eta_i + v_{it}$$

$$(t = q+1, \ldots, T_i \; ; \; i = 1, \ldots, N)$$

where η_i and λ_t are respectively individual and time specific effects, x_{it} is a vector of explanatory variables, $\beta(L)$ is a vector of associated polynomials in the lag operator and q is the maximum lag in the model. The number of time periods available on the ith individual, T_i, is small and the number of individuals, N, is large. Identification of the model requires restrictions on the serial correlation properties of the error term v_{it} and/or on the properties of the explanatory variables x_{it}. It is assumed that if the error term was originally autoregressive, the model has been transformed so that the coefficients α's and β's satisfy some set of common factor restrictions. Thus only white noise or MA errors are explicitly allowed. The v_{it} are assumed to be independently distributed across individuals with zero mean, but arbitrary forms of heteroskedasticity across units and time are possible. The x_{it} may or may not be correlated with the individual effects η_i and for each of these cases they may be strictly exogenous, predetermined or endogenous variables with respect to v_{it}.

The T_i equations for individual i can be conveniently written in the form:

$$y_i = W_i \delta + \iota_i \eta_i + v_i \quad ,$$

where δ is a parameter vector including the α_t's, the β's and the λ's, and W_i is a data matrix containing the time series of the lagged endogenous variables, the x's and the time dummies. Lastly, ι_i is a T_i x 1 vector of ones. DPD can be used to compute various linear GMM estimators of δ with the general form:

$$\hat{\delta} = \left[\left(\sum_i W_i^{*\prime} Z_i \right) A_N \left(\sum_i Z_i^{\prime} W_i^{*} \right) \right]^{-1} \left(\sum_i W_i^{*\prime} Z_i \right) A_N \left(\sum_i Z_i^{\prime} y_i^{*} \right)$$

where

$$A_N = \left(\frac{1}{N} \sum_i Z_i^{\prime} H_i Z_i \right)^{-1}$$

and W_i^* and y_i^* denote some transformation of W_i and y_i (e.g. first differences, orthogonal deviations, within groups, levels). Z_i is a matrix of instrumental variables which may or may not be entirely internal, and H_i is a possibly individual specific weighting matrix.

If the number of columns of Z_i equals that of W_i^*, A_N becomes irrelevant and $\hat{\delta}$ reduces to

$$\hat{\delta} = \left(\sum_i Z_i^{\prime} W_i^{*} \right)^{-1} \left(\sum_i Z_i^{\prime} y_i^{*} \right)$$

In particular, if $Z_i = W_i^*$ and the transformed W_i and y_i are deviations from individual means or orthogonal deviations[1], then $\hat{\delta}$ is the within-groups estimator. As another example, if the transformation denotes first differences, $Z_i = I_{T_i} \otimes x_i^{\prime}$ and $H_i = \hat{v}_i^* \hat{v}_i^{*\prime}$, where the v_i^* are some consistent first difference residuals, $\hat{\delta}$ is the generalised three stage least squares estimator of Chamberlain (1984). These two estimators require the x_{it} to be strictly exogenous for consistency. In addition, the within-groups estimator can only be consistent as $N \to \infty$ for fixed T if W_i^* does not contain lagged dependent variables and the model is truly static.

---

1 Orthogonal deviations as proposed by Arellano (1988) express each observation as the deviation from the average of future observations in the sample and weight each deviation to standardise the variance (i.e.
$x_{it}^* = [x_{it} - (x_{i(t+1)} + \ldots + x_{iT})/(T-t)](T-t)^{1/2}/(T-t+1)^{1/2}$ $t = 1,\ldots,T-1$)
If the original errors are iid, tranformed errors will also be iid.

When estimating dynamic models, we shall therefore typically be concerned with first difference or orthogonal deviations estimators that use different instrument sets at each time period, of the type discussed in Arellano and Bond (1988) and Arellano (1988). For example, if the panel is balanced, $p=1$, there are no explanatory variables nor time effects and the v_{it} are serially uncorrelated, then using first differences we have:

Equations	Instruments available
$\Delta y_{i3} = \alpha \Delta y_{i2} + \Delta v_{i3}$	y_{i1}
$\Delta y_{i4} = \alpha \Delta y_{i3} + \Delta v_{i4}$	y_{i1}, y_{i2}
.	.
.	.
$\Delta y_{iT} = \alpha \Delta y_{i(T-1)} + \Delta v_{iT}$	$y_{i1}, y_{i2}, \ldots, y_{i(T-2)}$

In this case $y_i^* = (\Delta y_{i3}, \ldots, \Delta y_{iT})'$, $W_i^* = (\Delta y_{i2}, \ldots, \Delta y_{i(T-1)})'$ and

$$Z_i = \begin{pmatrix} y_{i1} & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & y_{i1} & y_{i2} & \ldots & 0 & 0 & \ldots & 0 \\ . & . & . & & . & . & & . \\ . & . & . & & . & . & & . \\ 0 & 0 & 0 & \ldots & y_{i1} & y_{i2} & \ldots & y_{i(T-2)} \end{pmatrix}$$

Notice that precisely the same instrument set would be used to estimate the model in orthogonal deviations. Where the panel is unbalanced, for individuals with incomplete data the rows of Z_i corresponding to the missing equations are deleted, and missing values in the remaining rows are replaced by zeros.

In DPD we call one-step estimates those which use some known matrix as the choice for H_i. For a first-difference procedure, the one-step estimator uses

$$H_i = \begin{pmatrix} 2 & -1 & \ldots & & 0 \\ -1 & 2 & \ldots & & 0 \\ . & . & & & . \\ . & . & & & . \\ . & . & & & -1 \\ 0 & 0 & \ldots & -1 & 2 \end{pmatrix}$$

while for a levels or orthogonal deviations procedure the one-step estimator sets H_i to an identity matrix. If the v_{it} are heteroskedastic, a two-step estimator which uses

$$H_i = \hat{v}_i^* \hat{v}_i^{*\prime} \quad ,$$

where v_i^* are one-step residuals, is more efficent (cf. White (1982)). In models with explanatory variables, Z_i may consist of sub-matrices with the block diagonal form as above (exploiting all or part of the moment restrictions available), concatenated to straightforward one-column instruments. A judicious choice of the Z_i matrix should strike a compromise between prior knowledge (from economic theory and previous empirical work), the characteristics of the sample and computer limitations (see Arellano and Bond (1988) for an extended discussion and illustration). For example, if a predetermined regressor x_{it}, correlated with the individual effect, is added to the model discussed above, i.e.

$$E(x_{it} v_{is}) = 0 \quad \text{for} \quad s \geq t$$

$$\neq 0 \quad \text{otherwise}$$

$$E(x_{it} \eta_i) \neq 0 \quad ,$$

then the corresponding optimal Z_i matrix is given by

$$Z_i = \begin{pmatrix} y_{i1} & x_{i1} & x_{i2} & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & y_{i1} & y_{i2} & x_{i1} & x_{i2} & x_{i3} & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & \cdot & & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & y_{i1} & \cdots & y_{i(T-2)} & x_{i1} & \cdots & x_{i(T-1)} \end{pmatrix}$$

Where the number of columns in Z_i is too large, computational considerations may require those columns containing the least informative instruments to be deleted.

The assumption of lack of serial correlation of v_{it} is essential for the consistency of estimators such as those considered in the previous examples, which instrument the lagged dependent variable with further lags of the same variable. Thus DPD reports tests for the lack of first-order and second-order serial correlation in the residuals. If the model has been transformed to first differences, first-order serial correlation is to be expected but not second-order. If the model has been transformed to orthogonal deviations such first-order serial correlation is not induced (see

footnote 1). These tests are based on the standardised residual autocovariances which are asymptotically N(0, 1) variables under the null of no autocorrelation. More generally, Sargan tests of overidentifying restrictions are also reported. That is, if A_N has been chosen optimally for any given Z_i, the statistic

$$s = \left(\sum_i \hat{v}_i^{*\prime} Z_i \right) A_N \left(\sum_i Z_i^{\prime} \hat{v}_i^{*} \right)$$

is asymptotically distributed as a chi-square with as many degrees of freedom as overidentifying restrictions, under the null hypothesis of the validity of the instruments. Again, Arellano and Bond (1988) provides a complete discussion of these procedures.

3. USING DPD

All sample selection information and most model selection information are input to DPD by editing the DPD.RUN file. This can be done using the Gauss editor or any other compatible editor, and only a very basic knowledge of Gauss is needed to operate DPD. Options to include a constant and dummy variables, and selection from the menu of estimators, are controlled interactively when running DPD.

3.1 USER INPUT INFORMATION – THE DPD.RUN FILE

The DPD.RUN file is organised into several sections, each with a title and comments to provide "on-line" assistance. These sections are discussed in the same order here.

Data Set Selection

The data to be used in estimation are selected at the top of DPD.RUN. The main data set is specified by typing the name of the Gauss data set in the open statement at line 4. The auxiliary data set is selected in the same way at line 7. If the data sets are not on the same drive and subdirectory as DPD.RUN then their location must also be specified in these statements. It is strongly recommended that data is read from the hard disk only.

Immediately beneath each of these open statements, the variables **startf1** and **startf2** must be entered. These control the line in each data set at which DPD begins reading, and should always take positive integer values. If all the data is to be used then both these variables should be set to 1. However, if the main data set is sorted by some characteristic of the individual units, this feature will allow estimation on a sub-sample. For example, if in an unbalanced panel the data are sorted in ascending order of time-series observations per unit, then the balanced sub-panel with data for all time periods can be selected in this way. When using this facility, note that the data to be read starting from line **startf1** in the main data set must correspond to the data described from line **startf2** in the auxiliary data set. Modifying either **startf1** or **startf2** without the other is a recipe for trouble!

Other Data Information

The next section sets up several variables that are needed in reading the main data set and for creating dummy variables. All should take integer values. The first of these is **ncomp**, which controls how many cross-section units are read and processed at the same time. DPD operates by reading and processing the data in blocks of **ncomp** (or less) units. Execution speed increases with **ncomp**, but workspace constraints limit the number of units that can be analysed in each block. If an "insufficient workspace" or "read too large" error message is encountered, the solution is to reduce **ncomp**. The precise limit depends on the number of instruments in the model and (to a lesser extent) on the maximum number of observations per unit. As a rough guide, in samples like that used in the example below, we have found that when the number of instruments is around 20 then 15 or more units can be read in at the same time. As the number of instruments increases to 90 then only 2 or 3 units can be read in at once.

The next variable, **yearcol**, simply indicates the column of the main data set that contains the year to which each observation refers. One column must contain this information to allow testing for serial correlation, parameter stability and the creation of time dummies.

Cross-section units may also be associated by some observed characteristic, and if such a group indicator is available in the data then DPD will create intercept dummies according to this characteristic. With company data this will typically indicate an industry grouping to which the firm

belongs. This is assumed in what follows but need not be the case in practice. The variable indcol indicates the column of the main data set that contains the industry code to which each observation is classified. The variable indmax indicates the number of groups or classes that have been used. Where this option is used the industry codes in the data set should be integers running from one to indmax. Note that this is only an option. Where such a classification is either not desired or unavailable, indcol may be set to any arbitrary column number in the data set and indmax may be set to any arbitrary value. In this case the creation of "industry" dummies (see below) should not be requested.

The next variable, year1, indicates the year to which the earliest observations in the data set refer. This should be in the form 19xx, and is used to allocate time dummies to calendar years and for testing.

Finally in this section the variable lag must be entered. This controls how many time series observations on each unit are reserved to allow the creation of lagged series. DPD will automatically retain one observation for the creation of first differenced series, and if lag is not set to zero a further lag observations will be lost. Note that this effectively controls the maximum sample period that is available for estimation. Thus if year1 is 1971 and lag is 3, then estimation will use periods from 1975 onwards.

Data Transformations

The next section of DPD.RUN defines a Gauss subroutine in which data transformations and model selection are performed. Whenever this subroutine is called, all the columns of the main data set for the current block of units will have been read into a matrix called data. At this point, any operation that is available in Gauss may be performed on the columns of data in order to effect data transformations. Suppose for example that the data set contains 6 columns, and it is desired to use the ratio of the variables in columns 5 and 6 as a regressor in the model. This can be achieved by typing this pair of statements at the top of the subroutine:

```
temp = data[. , 5]./data[. , 6];
data = data~temp;
```

The first statement here picks out column 5 of data and divides each element by the corresponding element of column 6. The result is assigned to a vector called temp. The second statement then attaches this vector to the right hand side of data ("horizontal concatenation"). Thus data now has 7 columns, and the new variable occupies column 7. In this way transformed variables can be used in the model without being permanently stored in the data set. Any Gauss operations (e.g. logarithms, powers) can be performed similarly, and the variable name temp is reserved for this purpose. It is essential that after performing transformations the modified data matrix continues to have the name data, since DPD will look for this matrix when selecting the model.

Model Selection

The model to be estimated is selected using simple DPD functions. As in the econometric discussion above the dependent variable is y, the regressor matrix is x and the instrument matrix is z. The variables selected are also given names that will appear in the output.

a) Dependent variable

The dependent variable is selected with the functions lev(c,l), dif(c,l) or dev(c,l) which return a series in levels, first-differences or orthogonal deviations respectively. In all cases the first argument c indicates the column of data which contains the basic variable and the second argument l indicates the lag length to be produced. For example the statement

```
y = lev(3,0);
```

selects the variable in column 3 of data to be the dependent variable, in levels form. Similarly

```
y = dif(4,0);
```

selects the variable in column 4, and uses it in first-differenced form. Typically the lag length will be zero when selecting the dependent variable, although this is not essential.

After making y in this way, the selected variable must also be given a name. This is entered immediately below as the variable namey. The name has a maximum length of eight characters and must be enclosed between double inverted commas ("). Both upper and lower case may be used. For example, any of the following are acceptable:

namey = "v3";

or namey = "OUTPUT";

or namey = "log N";

b) Regressors

The same functions lev(c,l), dif(c,l) and dev(c,l) are used to select the matrix of regressors. Single columns are combined into a matrix using the horizontal concatenation operator (~) in Gauss. Different transformations may be combined, and any lag lengths up to the user-specified maximum (see above) are available. For example, the statement

x = dev(7,0)~dev(7,1);

selects a matrix of 2 regressors, both formed from the basic variable in column 7 of data and both in orthogonal deviations form. The first regressor is not lagged, and the second regressor is lagged one period.

Each of the regressors chosen must again be given a name. Names are also combined using horizontal concatenation, and each name must be enclosed in inverted commas. These are entered as the variable namex. For example, the statement

namex = "DQ"~"DQ(-1)";

could correspond to the regressor matrix specified above.

The only regressors that are not chosen in this way are the constant and intercept dummies. These may be added to the model interactively, and if they are selected then names are automatically assigned by DPD.

c) Instruments

A matrix of instruments may be formed using lev and dif in the same way as the matrix of regressors above. In addition DPD has a fourth function, gmm(c,l,n), which automatically returns all or part of the optimal instrument matrix required for the Generalised Method of Moments estimator discussed above. Again c indicates the column position of the basic variable in data. In a first difference model l refers to the lag length of the latest instrument to be exploited in each cross section. Thus if observations dated t-2 (and earlier) are taken to be valid instruments then l

would be set to 2. Since orthogonal deviations are forward differences "the equation for period t" in first differences corresponds to "the equation for period (t+1)" in orthogonal deviations, and with this proviso in mind the same instrument set would apply to both transformations. So with orthogonal deviations we may think of (l−1) as being the lag length of the most recent instrument. The third argument n indicates the (maximum) number of moment restrictions involving this basic variable to be exploited in each cross-section. A default value of 99 will exploit all available linear moment restrictions, as required for asymptotic efficiency. However, the current space limitations in Gauss restrict the instrument matrix to a maximum of 90 columns. Where this restriction is binding it may be preferred to exploit fewer moment restrictions for a larger number of basic variables. This can be achieved by choosing a suitably low value for n. For example, with first differences, if l is 2 and n is set to 3, then all columns of the optimal instrument matrix that do not contain instruments dated t−2, t−3 or t−4 are automatically discarded. This involves a loss in efficiency but ensures that the matrix returned has at most (3 x T^*) columns, where T^* is the number of cross-sections used in estimation. Unbalanced panel considerations are dealt with in the way discussed above.

Matrices produced by gmm may be combined with each other, or with vectors produced by lev or dif, again using horizontal concatenation. Note that the 90 column limit applies to the combined instrument matrix, including any dummy variables selected later. Where the chosen instrument set exceeds this maximum this is detected as soon as DPD is run, and an error message is returned.

The columns of x are not automatically included in the instrument matrix. When OLS rather than an instrumental variables estimator is required this is specified interactively from the menu of estimators. In this case the selection for z is arbitrary, although some contents must be entered.

In contrast to x, a name does not have to be specified for each column of z. A list of names which summarises the instrument set should however be entered as the variable namez. Again when the OLS option is used this selection is quite arbitrary, but some name must be entered.

We close this section with two examples that illustrate the syntax:

```
z = gmm(7,2,99);
namez = "Q(2,ALL)";
```

or

```
z = gmm(6,2,4)~gmm(7,2,3)~lev(3,2);
namez = "Y(2,4)"~"Q(2,3)"~"N(-2)";
```

Testing for Stability – the dbreak function

Having described the selection of variables for the model we are now in a position to describe the dbreak function, which allows tests of sub-sample parameter stability to be computed. The function dbreak(startyear,stopyear) returns a column vector containing a dummy variable which takes the value of 1 where observations relate to periods between startyear and stopyear (inclusive), and 0 elsewhere. The arguments are both entered in the form 19xx. When both arguments are the same the dummy takes the value 1 in that year only.

Stability tests are achieved by including as regressors both full-sample variables and the same variables multiplied by this dummy. For example, we may have

```
x = dif(7,0)~dif(7,1)~(dbreak(1980,1984).*(dif(7,0)~dif(7,1)));
```

The user-defined Wald test (see below) can then be used to test whether some or all of the estimated coefficients on the sub-sample variables are significantly different from zero. When using this option, ensure that namex contains a name for each of the columns in x (there are 4 columns in the above example).

The User-Defined Wald Test

When DPD is run it will automatically compute a Wald test of joint significance for all the variables entered in x (i.e. a test of the null hypothesis that their estimated coefficients are all zero). When intercept dummies are selected, similar tests of their joint significance are computed. In addition the user may select a subset of the regressors in x to be separately tested. This is useful in testing for sub-sample stability as well as more general linear restrictions.

This option is turned on by setting the variable waldtest to 1. Otherwise waldtest should be set to 0. When the option is selected, the columns of x that are to be tested are specified by

entering their column numbers as the variable testcols. These refer to column positions in x rather than data, and are combined using horizontal concatenation. Thus in the example of the previous sub-section, the stability test could be performed with the statement

```
testcols = 3~4;
```

As usual an arbitrary value should be assigned to testcols when this option is not being used.

Saving the Output

Output from DPD will appear on the screen but should also be directed to an output file for subsequent inspection and printing. This is accomplished by typing a filename in the output file statement at the bottom of DPD.RUN. Any DOS filename may be used here, and a location other than the default drive (including a floppy disk) may be specified. After the filename one of the words on or reset should appear. For example

```
output file = a:results.dat on;
```

If on is used the output from this run will be appended to the bottom of the output file. If reset is used the file will be overwritten. Care should be exercised when using the latter option!

3.2 RUNNING DPD

Once DPD.RUN has been edited the program is ready to run. Assuming that the three DPD files have been copied on to the \GAUSS subdirectory and that Gauss has been entered, then DPD can be run from command mode with the command

```
run dpd.run
```

If DPD.RUN is not on the \GAUSS subdirectory then its location must also be specified in the run command. Alternatively DPD can be run from the edit mode, which is often most convenient. Enter the Gauss editor with the command

```
edit dpd.run
```

and use the F2 key from the editor to execute the program.

3.3 THE MENU OF OPTIONS

On running DPD the user is presented with a series of options which are controlled by typing answers to prompts on the screen. The first question asks for the form of the model to be entered. This is achieved by typing 0 if a levels model is required, 1 if a first-differenced model is required and 2 if an orthogonal deviations model is required. In each case the number entered should be followed by the return key. This information determines the form of the H matrix used to compute the one-step estimator as discussed in Section 2.

The second question asks for the estimation method to be used. Here typing 0 will produce ordinary least squares estimates and typing 1 will produce instrumental variables estimates using the instrument set selected in DPD.RUN.

The third question asks whether standard errors and test statistics that are consistent in the presence of general heteroskedasticity are to be computed. In this and the following questions the user should type 1 if this option is desired and 0 otherwise. Where appropriate the two-step instrumental variables estimator (see above) is also produced when this option is requested. This option is not automatic as it requires the data set to be read a second time and so increases the execution time of the program. However in our experience heteroskedasticity is often present in panel data models. In this case non-robust test statistics may be misleading and the two-step estimation procedure may offer significant improvements in precision, so that this option is strongly recommended.

The next set of questions determine whether a constant and intercept dummies are included in the model. Again DPD prompts with simple yes/no questions. Where intercept dummies (time or "industry") are selected the equation will include one less dummy variable than the total number of periods or classes available, together with a constant term, which is equivalent to the inclusion of a complete set of intercept dummies. For time dummies, the coefficient on the constant gives the intercept coefficient for the first cross-section used in estimation, and those on the dummy variables are deviations from this initial intercept value. The same principle applies with "industry" dummies, where again the dummy for class 1 is excluded. The Wald tests provided test the joint significance

of these dummy variables. With levels estimators the constant is excluded from the coefficients tested, but not for first-difference or orthogonal deviations estimators. Note that when the constant and dummies are selected as regressors, they are automatically added to the set of instruments used for IV estimation. However, the option of using the dummy variables as instruments but not as regressors is also available. Where dummies are not required either as regressors or as instruments the user must respond with 0 to both questions. In this case a constant intercept may be selected.

The last two questions ask respectively whether basic descriptive statistics and the covariance matrices for the estimated coefficients should be included in the output file. The descriptive statistics available show the mean, standard deviation and extreme values of each series in y and x, together with a matrix of simple correlation coefficients. Heteroskedasticity-robust covariance matrices for the one-step and two-step estimators are provided where these are appropriate.

3.4 OUTPUT FROM DPD

The output file produced by DPD is largely self-explanatory, and an example is included in the next section. The last column in the main tables, labelled P-Value, reports the probability that each coefficient may be zero, assuming that the errors are normally distributed. With the basic one-step estimates the residual sum of squares (RSS) and total sum of squares (TSS) are reported, along with the estimated variance of the error term in the levels model (even where a first-difference *estimator* is used). Note that this statistic is of less interest where heteroskedasticity is suspected.

The Wald tests reported are asymptotically distributed as χ^2 variables, with the degrees of freedom (df) provided. The Sargan tests of overidentifying restrictions are also asymptotically disributed as χ^2. The tests for first-order and second-order serial correlation relate to the residuals from the estimated equation, so that first-differencing will induce MA(1) serial correlation even where the error term in levels is a white-noise disturbance. These tests are presented in Arellano

and Bond (1988), and are asymptotically distributed as standard normal variables. Where robust test statistics are selected and the data is read a second time, the complete serial correlation matrix (based on the one-step residuals) is also computed.

4. AN EXAMPLE

In this section we present an example DPD.RUN file together with the output file that it produced. The example data sets XDATA and AUXDATA are supplied with DPD. XDATA has six columns which contain data for the sample of 140 UK quoted companies over the period 1976–1984 used in Arellano and Bond (1988). The variables in these columns are an industry code, the accounting year, employment, real wages, gross capital stock and an index of industry output respectively. The panel is unbalanced, with observations varying between 7 and 9 records per company.

The example DPD.RUN file specifies a log-linear labour demand equation including 2 lags of the dependent variable, current and lagged real wages, current capital and current and lagged industry output. Notice that the log series are constructed internally at the data transformations stage. The model is estimated in first differences. The instrument set exploits all available linear moment restrictions involving the dependent variable (assuming white-noise errors in levels), in combination with the remaining regressors in stacked form. A Wald test of the joint significance of the two real wage variables is computed.

On running DPD time dummies were requested, but not industry dummies. Robust test statistics and two-step estimates were selected. Descriptive statistics and covariance matrices of the estimates were omitted. This results in the output file reproduced here.

Example DPD.RUN file

```
new ,24000; #include dpd.fns;
@------------------------------ SPECIFY DATA SET ---------------------------

@ Specify main GAUSS data set @          open f1=xdata;
@ Start reading at line @                startf1=1;

@ Specify auxiliary data set @           open f2=auxdata;
@ Start reading at line @                startf2=1;

@------------------------------ DATA INFORMATION ---------------------------

@ Number of companies to @
@ process in each read @                 ncomp=10;

@ Data column for year @                 yearcol=2;
@ Data column for industry @             indcol=1;
@ Number of industry classes @           indmax=9;

@ First year of data @                   year1=1976;
@ Longest lag to be constructed @        lag=2;

@------------ DATA TRANSFORMATIONS AND MODEL SELECTION SUBROUTINE -----------
goto below;
model:

@ Data set is read in to the matrix data @
@ Operations using columns of data may be performed here @

data=ln(data);

@ SELECT DEPENDENT VARIABLE @

y=dif(3,0);
namey="Dn";

@ SELECT REGRESSORS @

x=dif(3,1)~dif(3,2)~dif(4,0)~dif(4,1)~dif(5,0)~dif(6,0)~dif(6,1);
namex="Dn(-1)"~"Dn(-2)"~"Dw"~"Dw(-1)"~"Dk"~"Dys"~"Dys(-1)";

@ SELECT INSTRUMENTS @

z=gmm(3,2,99)~x[.,3:7];
namez="n(2,all)"~namex[.,3:7];

@ Use dbreak(19xx,19xx) to select a sub-period @

return;
below:
@--------------- USER-DEFINED WALD TEST OF JOINT SIGNIFICANCE --------------

@ Set waldtest=1 for this option @       waldtest=1;
@ Select columns of x to be tested @     testcols=3-4;


@-------------------------- SPECIFY FILE FOR OUTPUT ----------------------

output file=xdata.out on;

@ Note that "on" will append and "reset" will overwrite @

@*****************************************************************************
@
@              AFTER SELECTING THE MODEL THIS PROGRAM CAN BE
@              RUN DIRECTLY FROM THE GAUSS EDITOR BY HITTING F2
@
@*****************************************************************************

#include dpd.prg; end;
@------------------------------ END OF PROGRAM -----------------------------
```

Example OUTPUT file

D.P.D. RESULTS

FIRST DIFFERENCES IV

Number of firms: 140 Sample period is 1979 to 1984
Observations: 611 Degrees of freedom: 598

Dependent variable is: Dn

Instruments used are:
CONST n(2,all) Dw Dw(-1) Dk Dys Dys(-1) TIM DUMS

ONE-STEP ESTIMATES

RSS = 8.219380 TSS = 12.599978
Estimated sigma-squared (levels) = 0.006872

Wald test of joint significance:	352.585004	df =	7
Wald test - jt sig of time dums:	11.254987	df =	6
Wald test selected by user:	91.830846	df =	2
Testing: Dw Dw(-1)			
Sargan test:	73.858107	df =	25

Var	Coef	Std. Error	T-Stat	P-Value
CONST	0.005427	0.012814	0.423552	0.671892
Dn(-1)	0.534614	0.127418	4.195740	0.000027
Dn(-2)	-0.075069	0.043441	-1.728079	0.083974
Dw	-0.591573	0.061907	-9.555793	0.000000
Dw(-1)	0.291510	0.095558	3.050603	0.002284
Dk	0.358502	0.034868	10.281733	0.000000
Dys	0.597199	0.127326	4.690304	0.000003
Dys(-1)	-0.611705	0.167947	-3.642250	0.000270
D80	0.005608	0.020075	0.279335	0.779988
D81	-0.038305	0.017635	-2.172098	0.029848
D82	-0.027785	0.018522	-1.500107	0.133587
D83	-0.006850	0.019021	-0.360148	0.718737
D84	0.006314	0.023754	0.265799	0.790394

NOTE: Standard errors and test statistics not robust to heteroskedasticity

Test for first-order serial correlation: -3.409
Test for second-order serial correlation: -0.369

ONE-STEP ESTIMATES WITH ROBUST TEST STATISTICS

Wald test of joint significance:	219.623310	df =	7
Wald test - jt sig of time dums:	11.450408	df =	6
Wald test selected by user:	12.486527	df =	2
Testing: Dw Dw(-1)			

Var	Coef	Std. Error	T-Stat	P-Value
CONST	0.005427	0.009714	0.558696	0.576369
Dn(-1)	0.534614	0.166449	3.211871	0.001319
Dn(-2)	-0.075069	0.067979	-1.104302	0.269462
Dw	-0.591573	0.167884	-3.523705	0.000426
Dw(-1)	0.291510	0.141058	2.066597	0.038772
Dk	0.358502	0.053828	6.660098	0.000000
Dys	0.597199	0.171933	3.473441	0.000514
Dys(-1)	-0.611705	0.211796	-2.888179	0.003875

D80	0.005608	0.015378	0.364660	0.715365
D81	-0.038305	0.017445	-2.195731	0.028111
D82	-0.027785	0.017908	-1.551541	0.120772
D83	-0.006850	0.022055	-0.310593	0.756110
D84	0.006314	0.019713	0.320284	0.748753

Robust test for first-order serial correlation: -2.493
Robust test for second-order serial correlation: -0.359

Estimated serial correlation matrix

```
 1.000
-0.478  1.000
 0.047 -0.141  1.000
 0.136 -0.170 -0.439  1.000
 0.259 -0.124  0.015 -0.454  1.000
 0.158 -0.190  0.050  0.412 -0.559  1.000
```

Number of observations available to sample covariances

```
 80
 80   138
 80   138   140
 80   138   140   140
 18    76    78    78    78
 14    33    35    35    35    35
```

TWO-STEP ESTIMATES

Wald test of joint significance: 371.987810 df = 7
Wald test - jt sig of time dums: 26.904500 df = 6
Wald test selected by user: 111.105724 df = 2
Testing: Dw Dw(-1)
Sargan test: 30.112471 df = 25

Var	Coef	Std. Error	T-Stat	P-Value
CONST	0.010509	0.007251	1.449224	0.147275
Dn(-1)	0.474151	0.085303	5.558424	0.000000
Dn(-2)	-0.052968	0.027284	-1.941316	0.052220
Dw	-0.513205	0.049345	-10.400258	0.000000
Dw(-1)	0.224640	0.080063	2.805799	0.005019
Dk	0.292723	0.039463	7.417736	0.000000
Dys	0.609775	0.108524	5.618817	0.000000
Dys(-1)	-0.446373	0.124815	-3.576284	0.000349
D80	0.003633	0.012734	0.285327	0.775394
D81	-0.050962	0.013710	-3.717118	0.000202
D82	-0.032149	0.013986	-2.298604	0.021527
D83	-0.012356	0.012842	-0.962161	0.335969
D84	-0.020730	0.013679	-1.515434	0.129662

Robust test for first-order serial correlation: -2.826
Robust test for second-order serial correlation: -0.327

Execution time is 397.220 seconds

REFERENCES

ARELLANO, M. (1988), "An alternative transformation for fixed effects models with predetermined variables", mimeo, Institute of Economics and Statistics, Oxford.

ARELLANO, M. and BOND, S.R. (1988), "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations", Institute for Fiscal Studies, Working Paper 88/4.

CHAMBERLAIN, G. (1984), "Panel Data", in Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, Volume II, Elsevier Science Publications.

WHITE, H. (1982), "Instrumental Variables Regression with Independent Observations", *Econometrica*, 50, 483–499.